Foodservice Recommender System

Springboard Data Science

Capstone project 2

Hye Joo Han

Project Goal

- Imaginary client: a food delivery service company

- Goal:
  - Analyze ratings for foodservice businesses
    (e.g., restaurants, coffee shops and breweries)
  - Build a recommender system models that predict ratings
  - Make business recommendations

Procedures

- **Data collecting and wrangling**
- Exploratory data analysis (EDA)
  - Businesses
  - Reviews
  - Users
- Recommender system models
- Final recommendation

Datasets

- Yelp datasets downloaded from https://www.yelp.com/dataset
- 3 files in json format.
  - review.json

    (review data with star rating, user_id, business_id, review date and comment)
  - business.json

    (business information in 10 metropolitan areas such as location data, number of reviews, average stars, attributes, and categories)
  - user.json

    (user information such as user_id, average stars and first date on Yelp)
- Cleaned each dataset (see the following slides)

Data Wrangling business.json

- 'categories' column
  - One top category and several subcategories related to the business
  - Selected only 'food' and 'restaurants’ top categories
- 72,624 'food' and 'restaurants’ businesses left
- Originally 18 columns, but 39 more added after expanding a dictionary column ‘attributes’
- Missing values:
  - Dropped the columns with unnecessary information or too many missing values
  - Missing city names were guessed using latitude and longitude
- No suspicious outliers

Data Wrangling
review.json

- String columns were cleaned
- Missing values
  - -1 for 'cool' or 'useful’ vote counts
  - 0 star (only 1-5 stars are possible)
  - Removed 3 rows with these missing values
- Selected only the reviews for foodservice businesses
- 4,017,884 reviews left after cleaning and filtering

Data Wrangling
user.json

- Many empty strings for user names and many ‘None’ (string type, not None type) for 'elite' and 'friends' columns. I did not clean these for now.
- I filtered out the users who have not left any reviews for foodservice businesses.
- 1,073,581 users left

Procedures

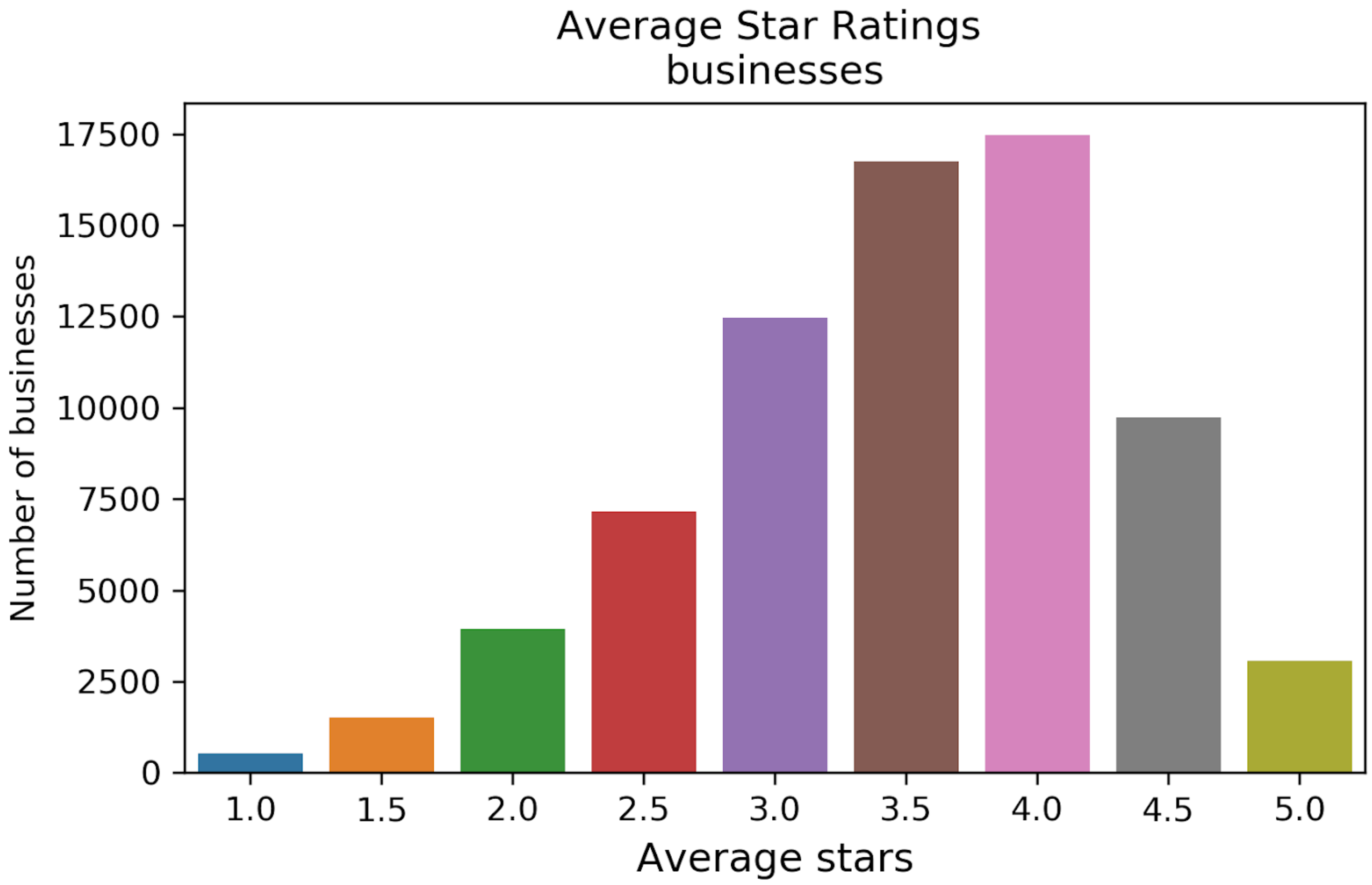
- Data collecting and wrangling
- **Exploratory data analysis (EDA)**
  - **Businesses**
  - **Reviews**
  - **Users**
- Recommender system models
- Final recommendation

Businesses

Where are the businesses?

- Mostly in North America and Europe
- Business in Antarctica? No, it was there due to swapped latitude and longitude

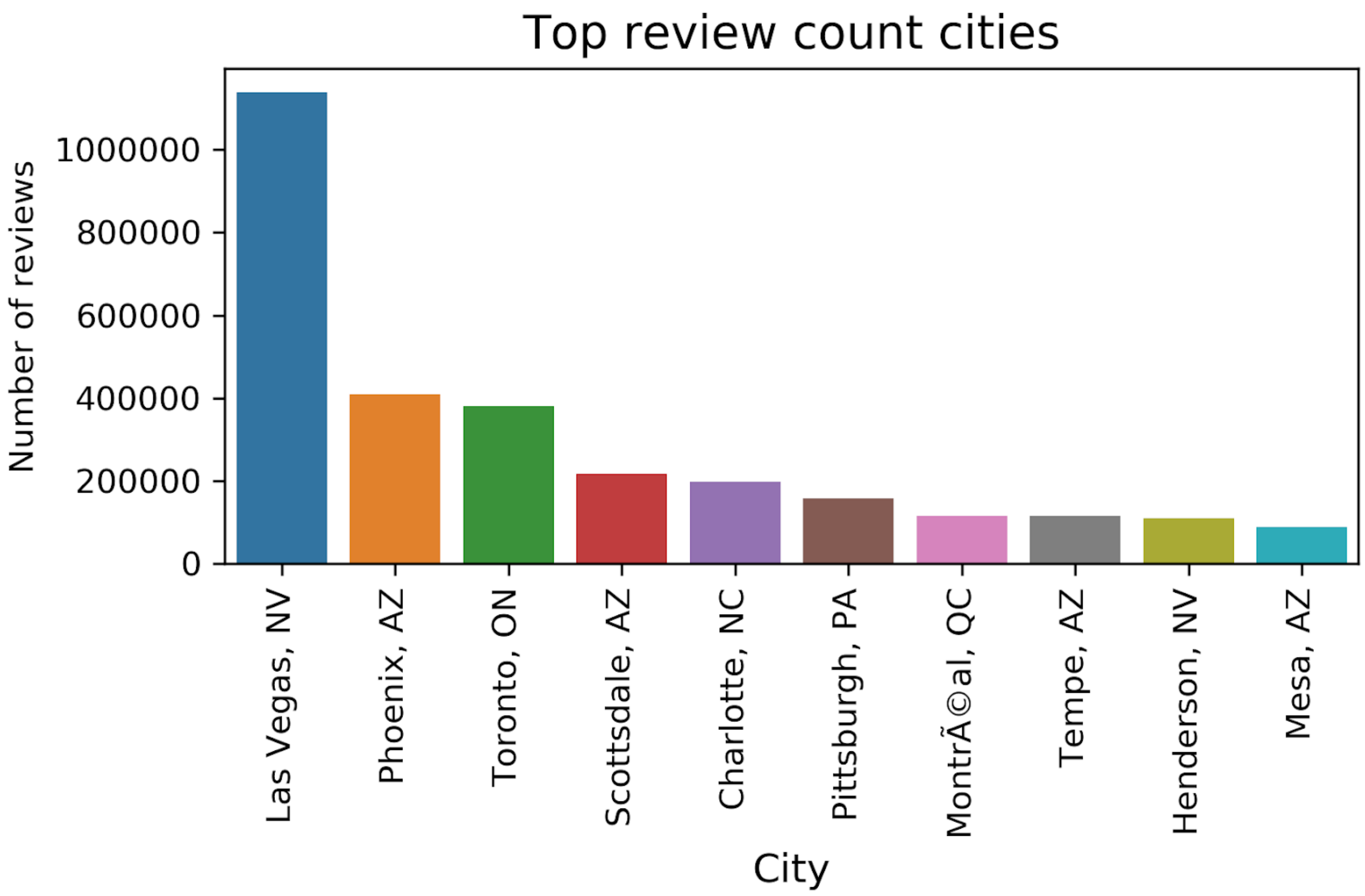
Average stars for business



Review count for business

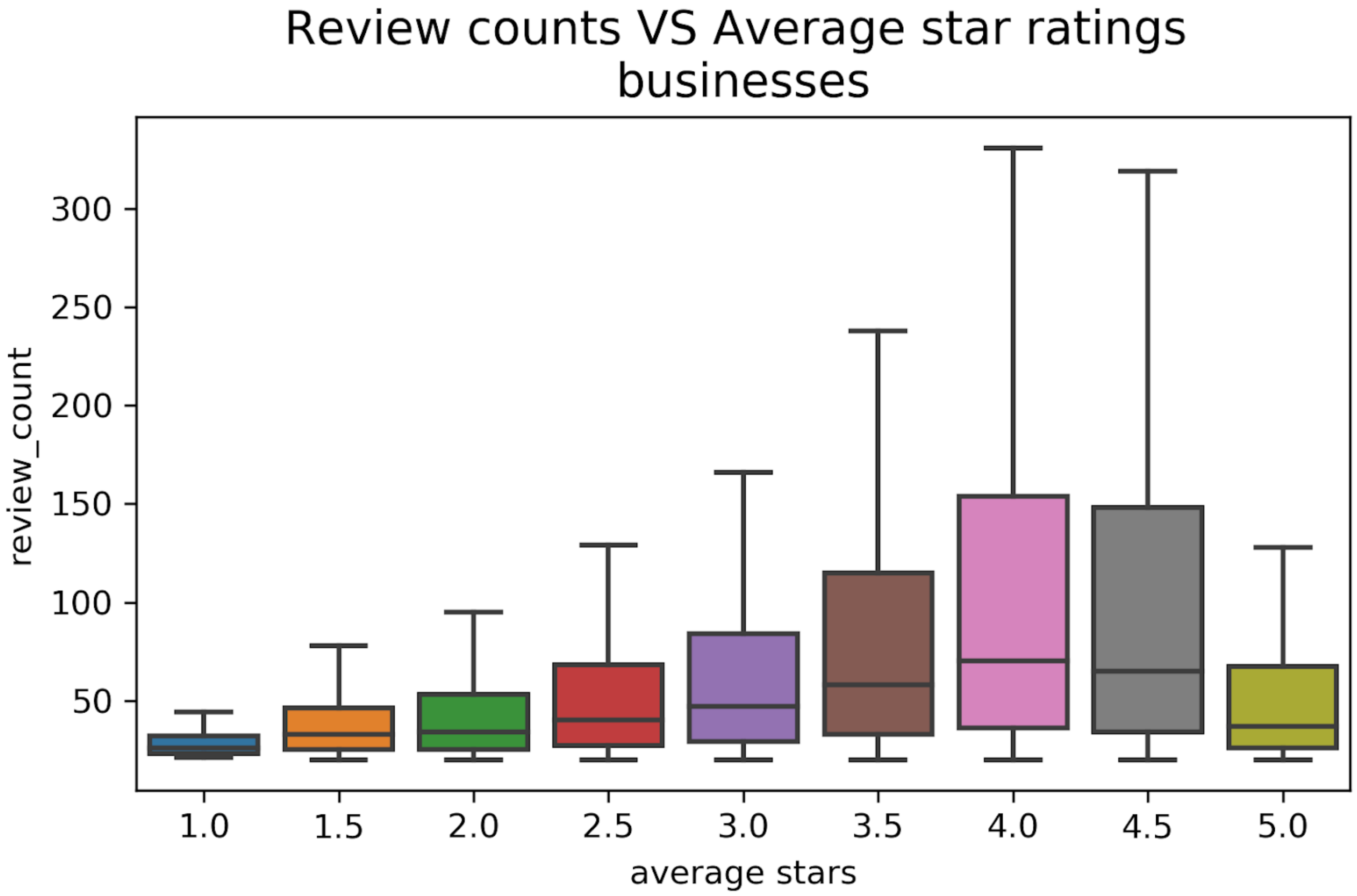
- Review counts for businesses are highly right-skewed
- Top 6 businesses with the highest review counts have
  - Locations in Las Vegas, NV
  - Average stars higher than the mean stars (3.49)
  - Price ranges higher than the average price range, 1.7, except for one
- 3 questions made from the above result:
  - Which cities have the most reviews
  - Whether highly rated businesses tend to get more reviews
  - Whether more expensive businesses tend to get more reviews

Cities with the most reviews



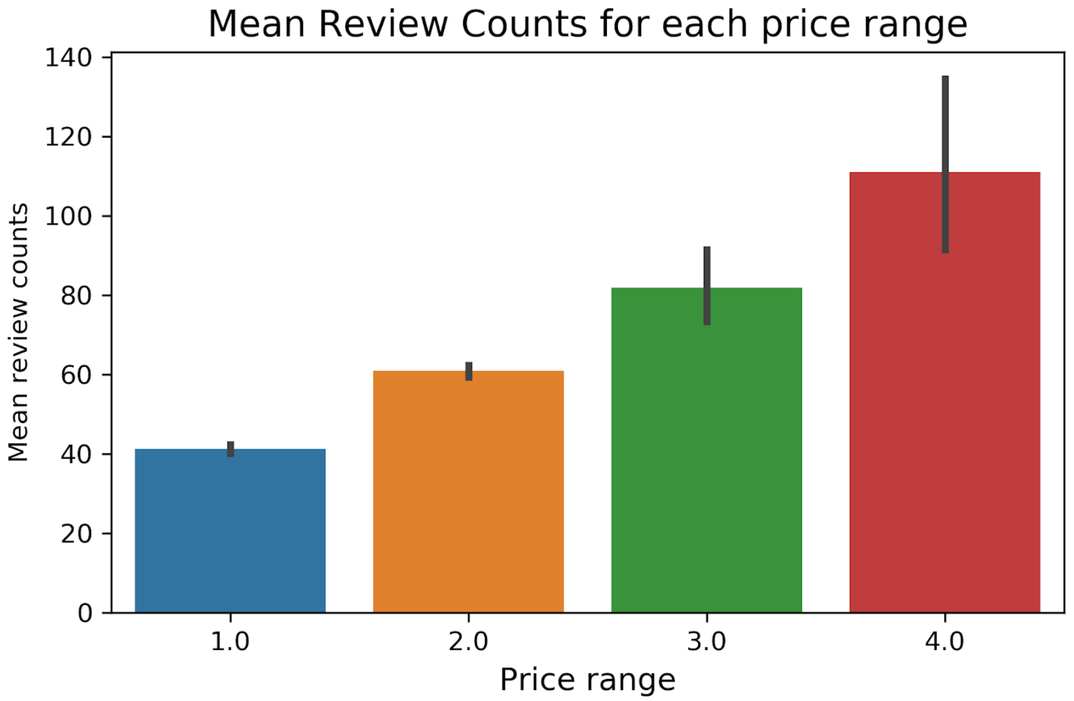
- The city with the most number of reviews is Las Vegas, NV.
- The 9th city, Henderson is also part of the Las Vegas metropolitan area.

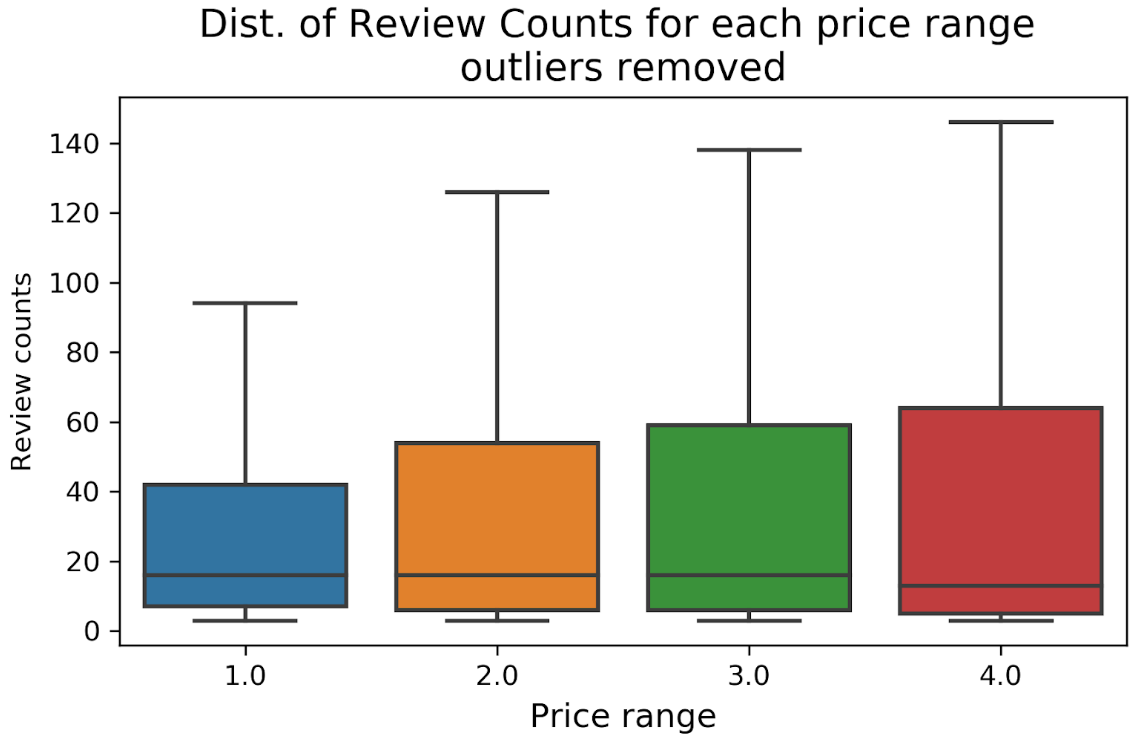
Do highly rated businesses tend to have more reviews?



- Highly rated businesses tend to have more reviews with some exceptions

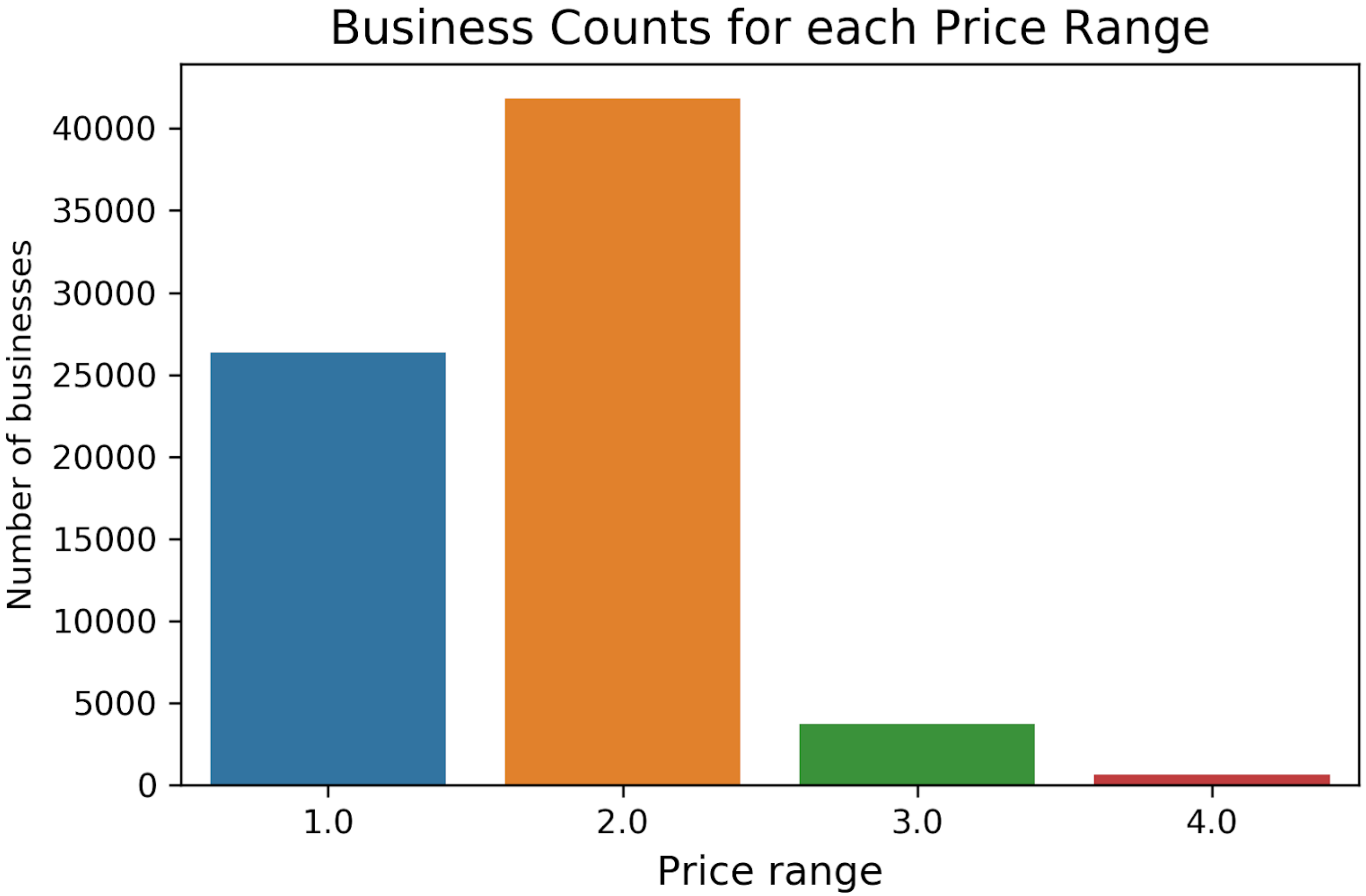
Do more expensive businesses tend to get more reviews?





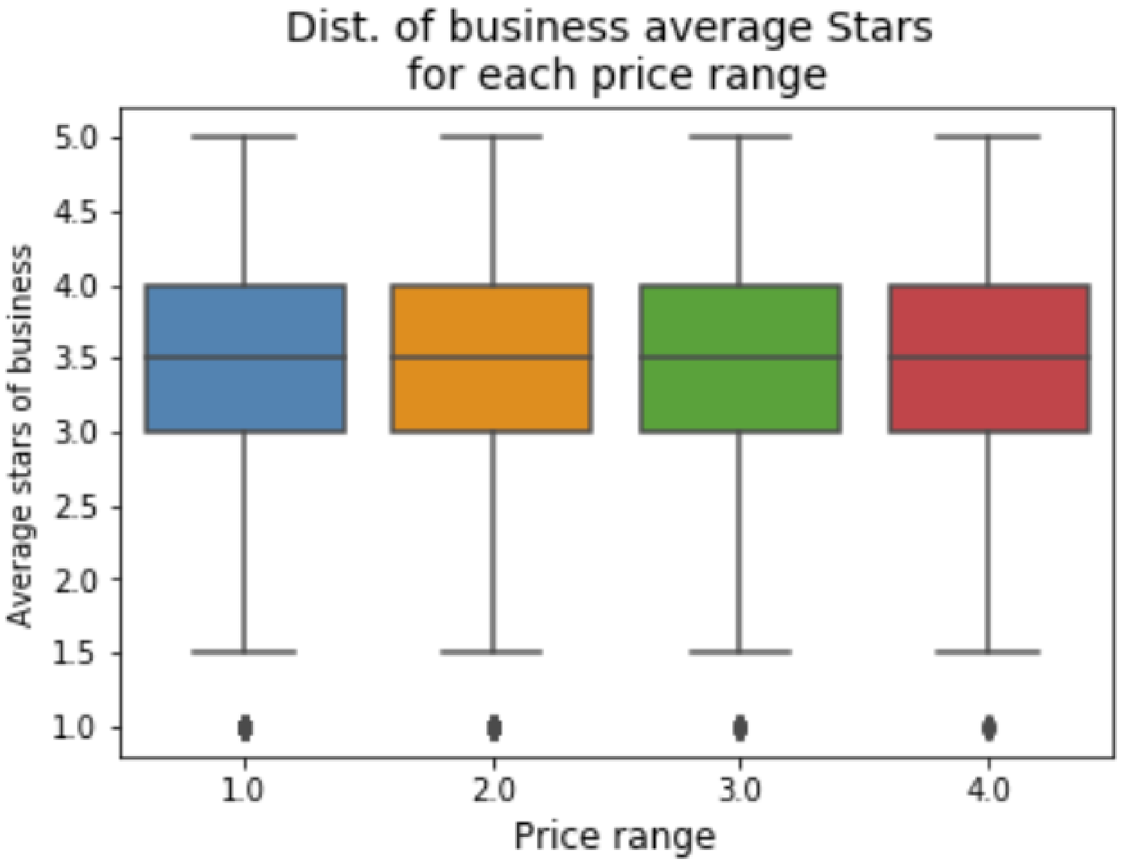
- The barplot (left) seems to show people are more likely to leave reviews for more expensive foodservice businesses
- However, the boxplot (right) shows there are not much difference in medians of review counts among the 4 different price ranges (outliers removed).

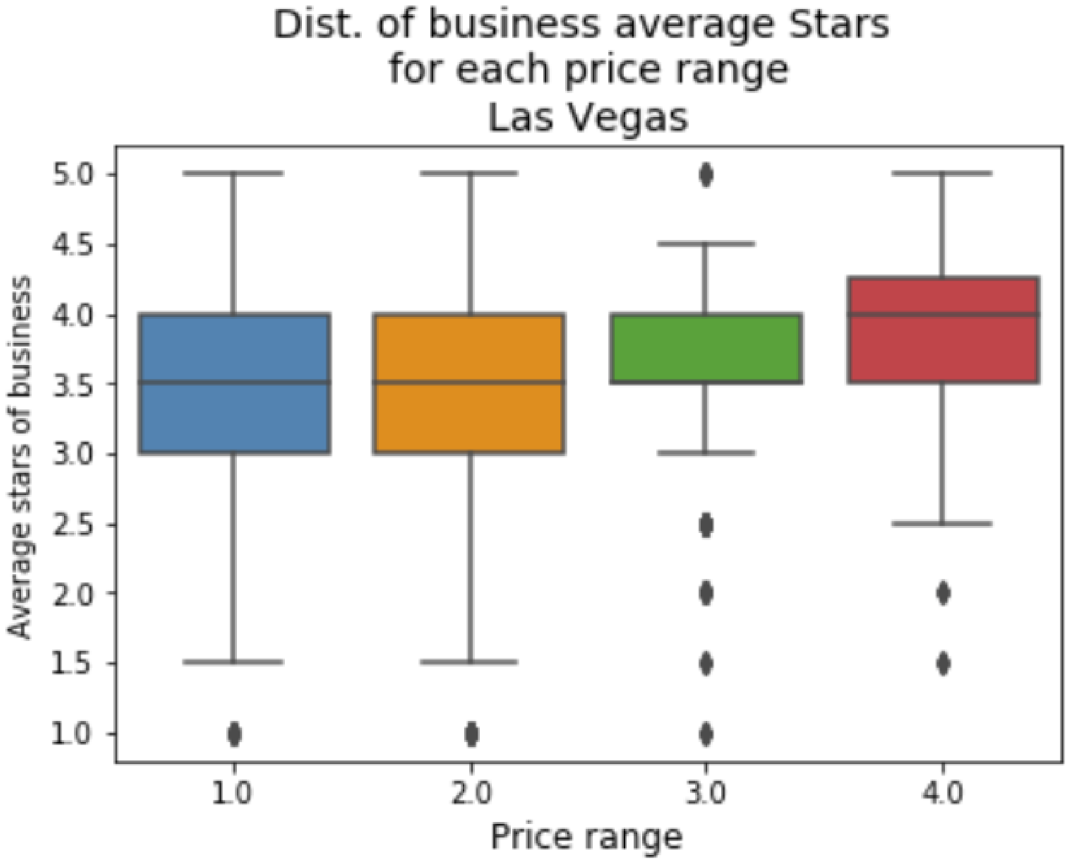
Price Range



- The most common price range is 2 and then 1
- Much fewer businesses with price range 3 and 4

Do more expensive businesses receive higher star ratings?

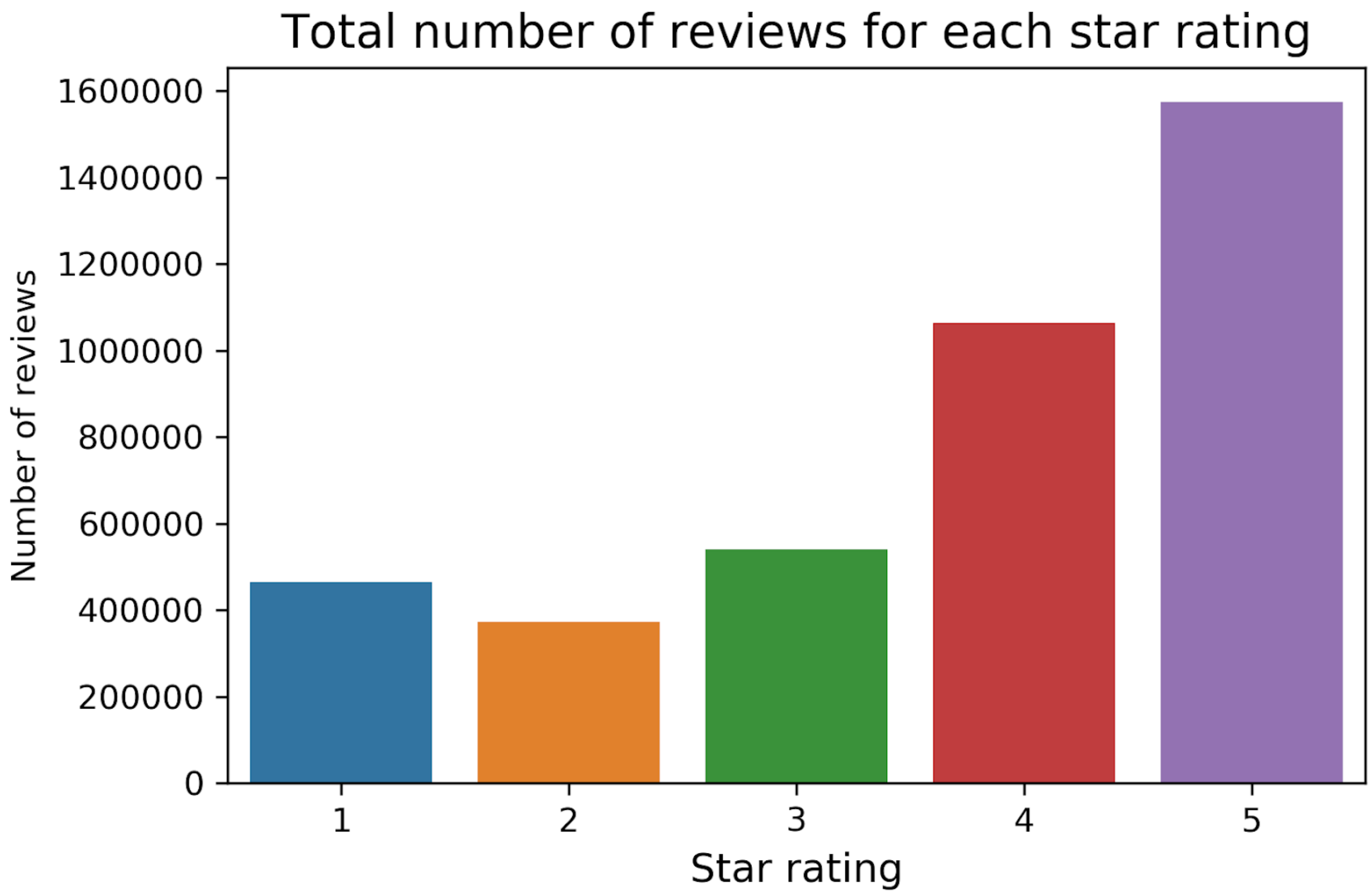




- No (left), but it could be true for some cities (right)

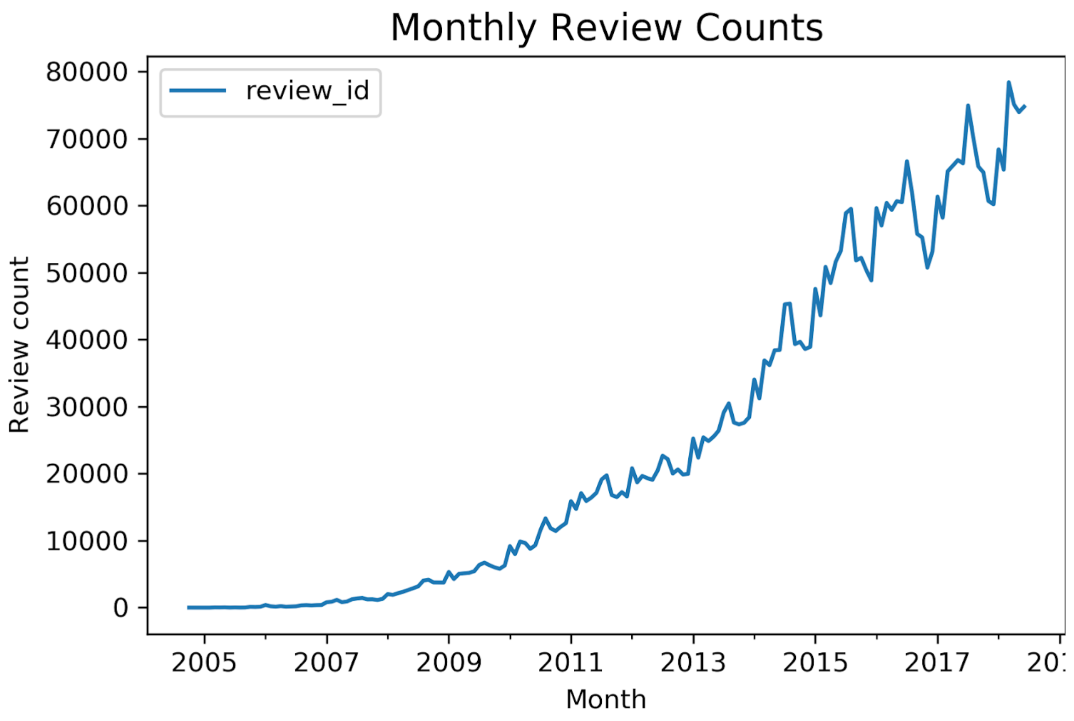
Reviews

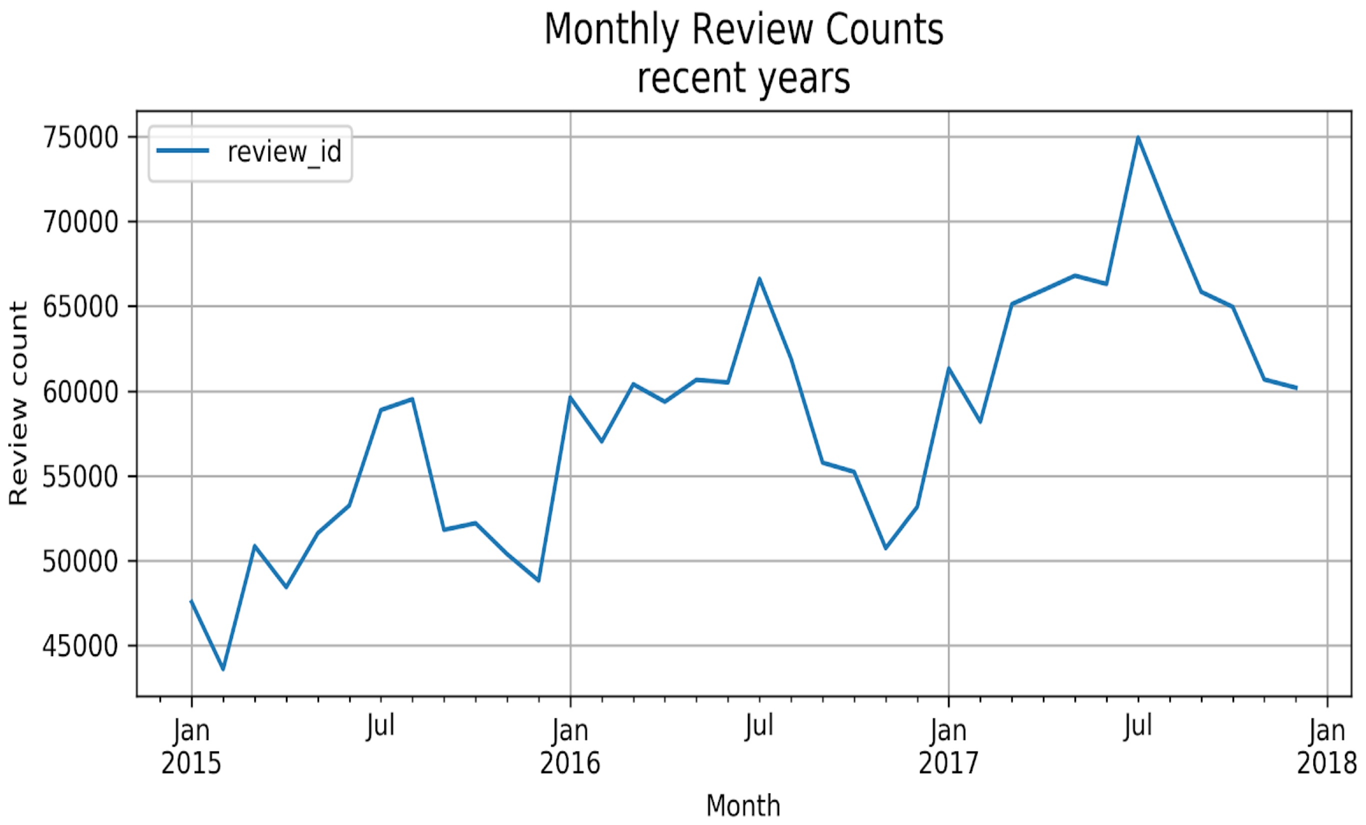
Frequency of each star in reviews



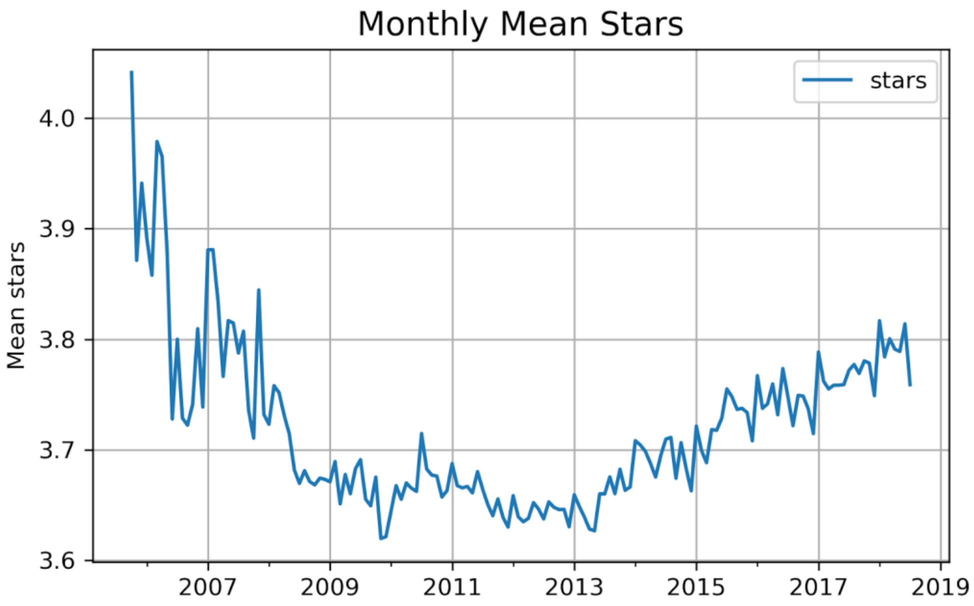
- Higher stars are more frequent except that 1 star is more frequent than 2 stars

Review date

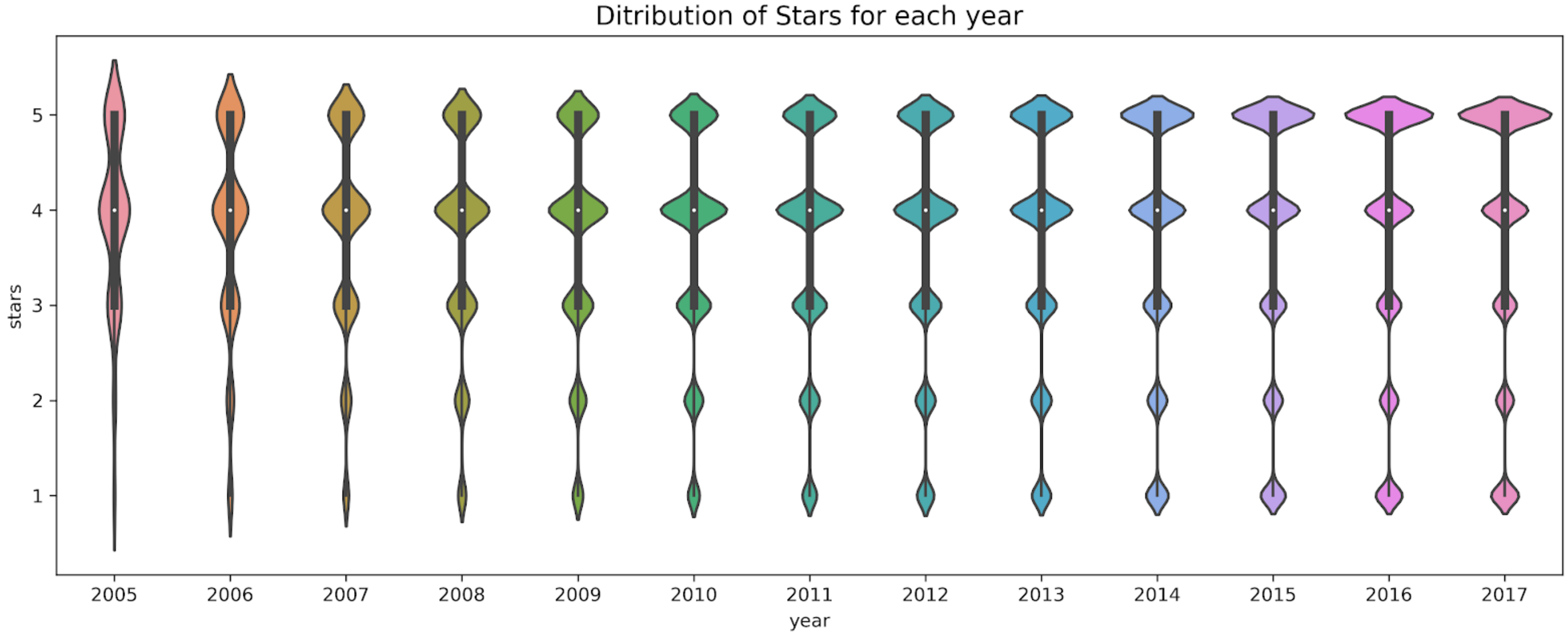




- Monthly number of reviews have exponentially increased over time
- The monthly review count drops during winter months and reaches seasonal peaks around July



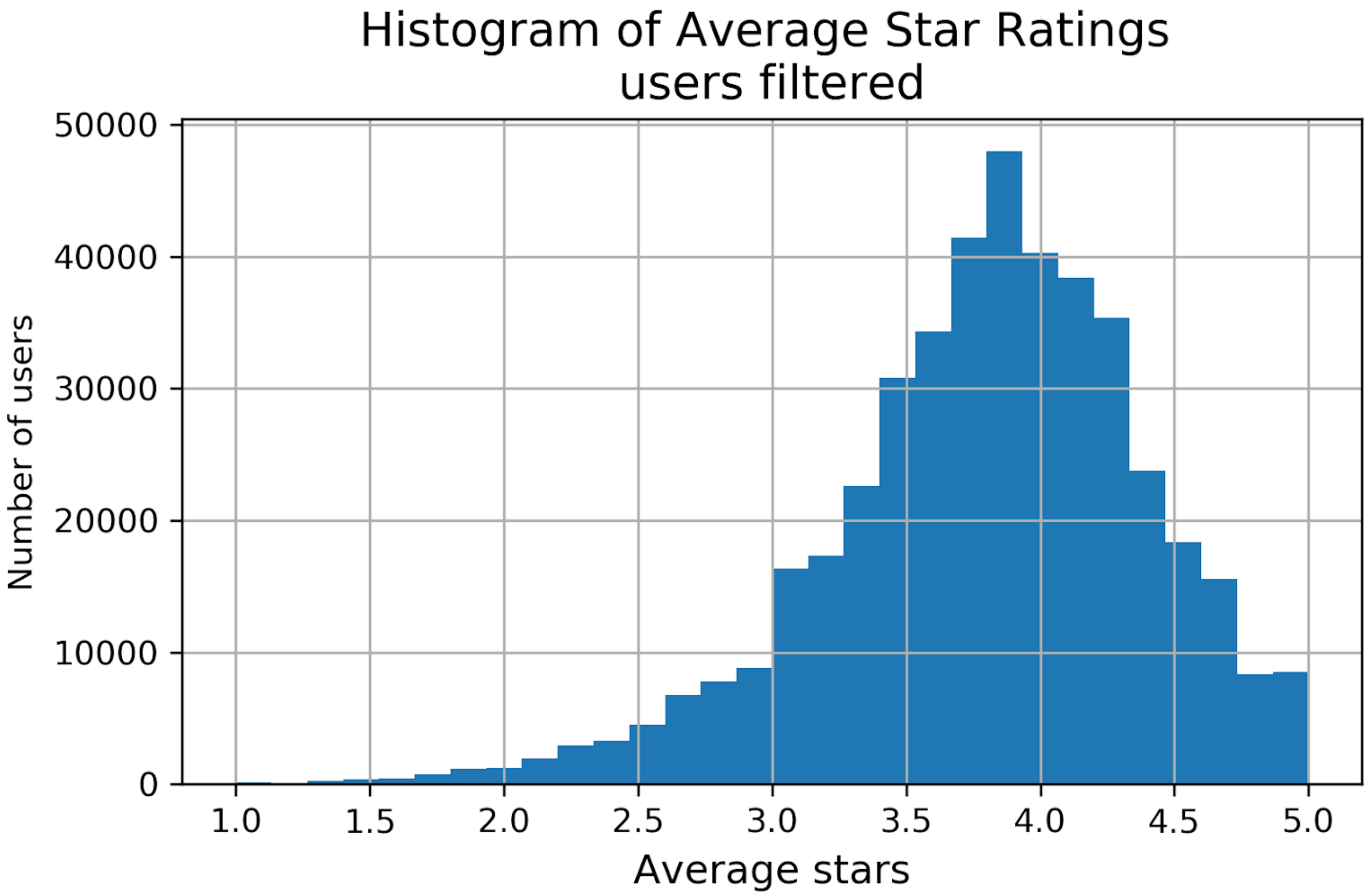
- The below violin plots show how distributions of stars changed over time and explain the quadratic shape in the graph for monthly mean stars (left).
- In the beginning, low stars were very rare and most stars were 3, 4, or 5. As years go by, 1 or 2 stars also became frequent. This can explain why the average stars were higher in the beginning and decreased over time.



- Up to 2013, 4 stars are the most frequent star rating, but from 2014 5 stars become the most frequent rating; this can explain the increase of average stars from 2014.

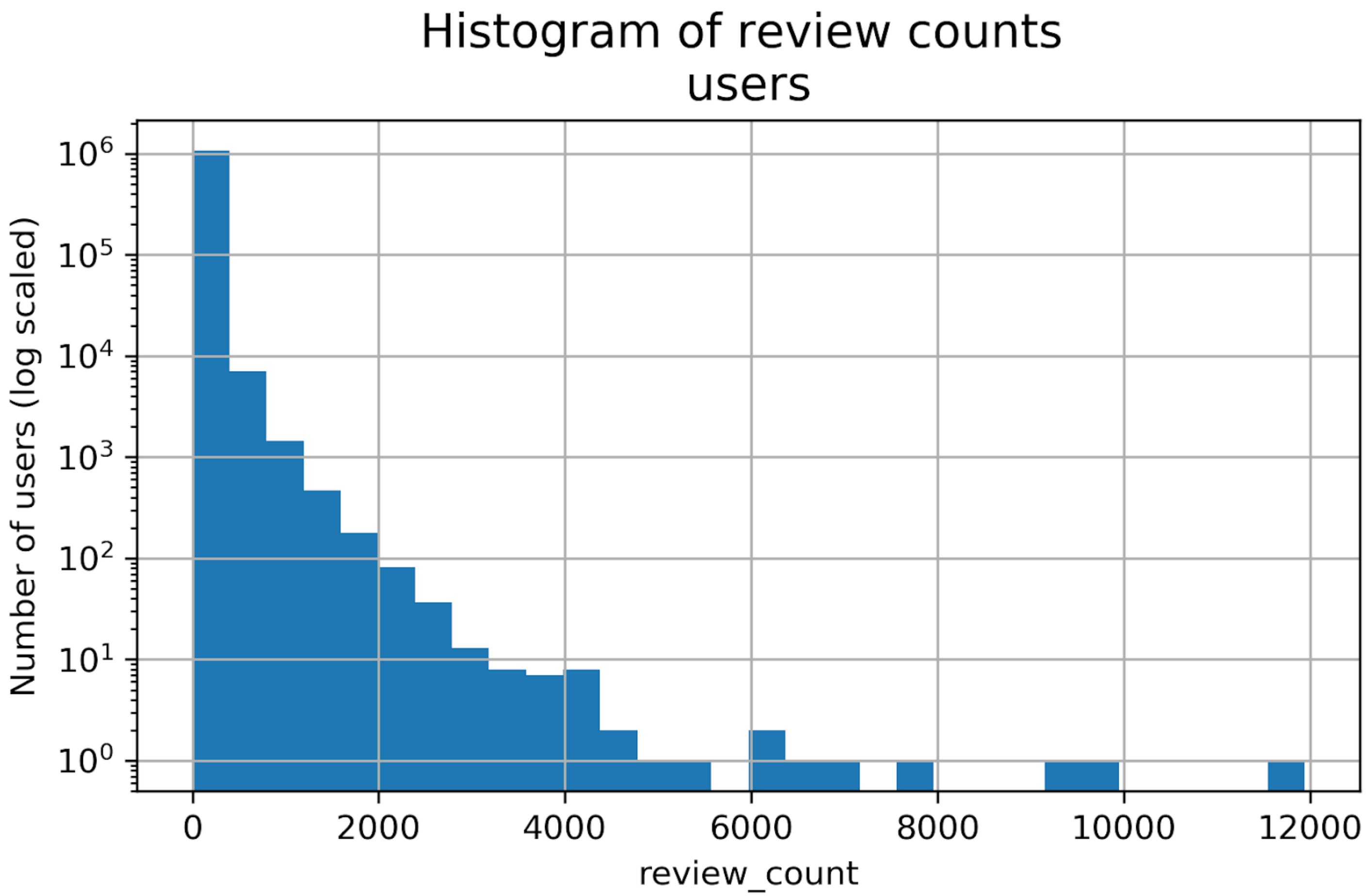
Users

Average stars for users



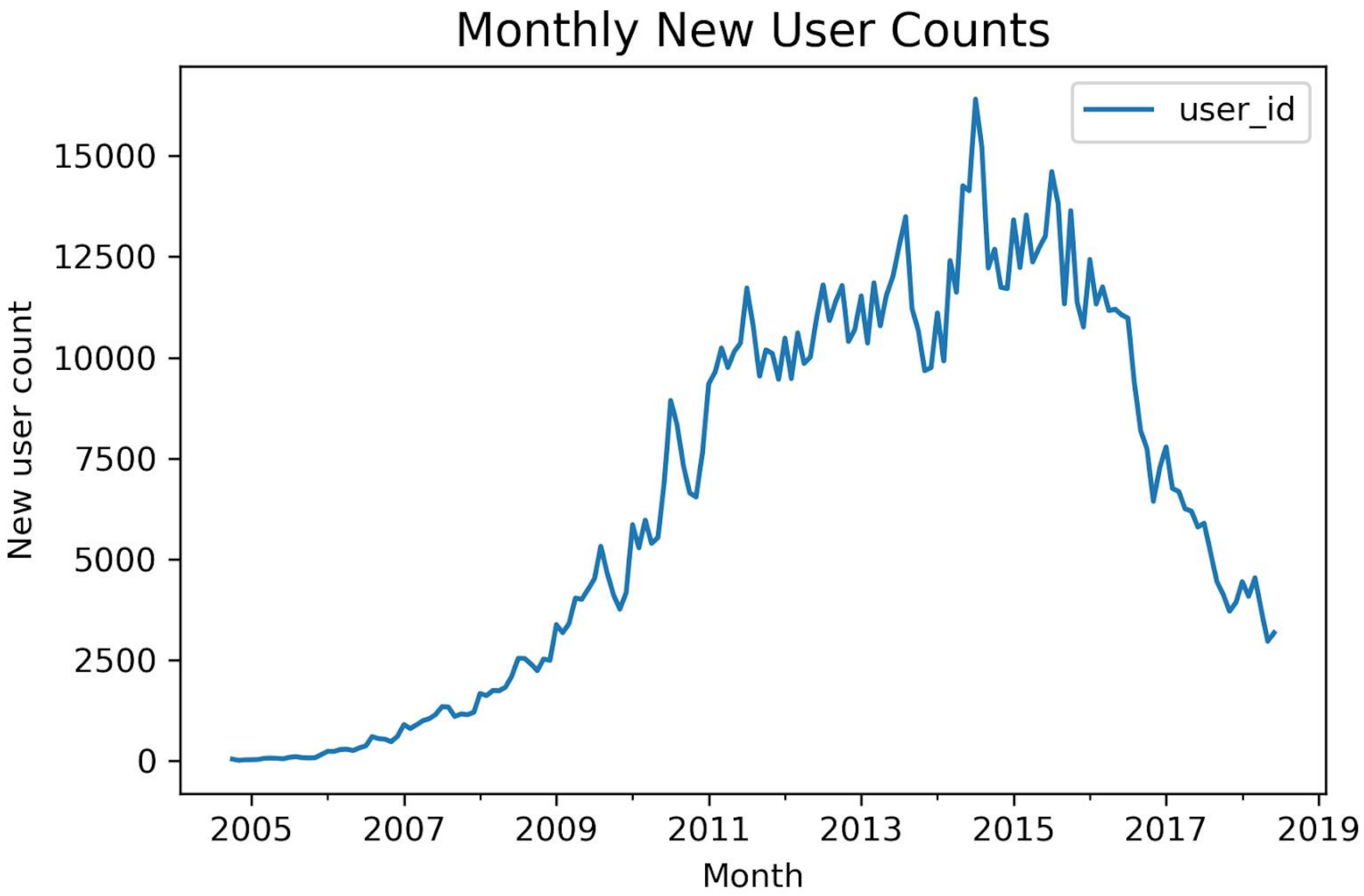
- Unimodal with a peak around 3.8 and is left-skewed
- Similar to the distribution of business average stars

Review Count for users



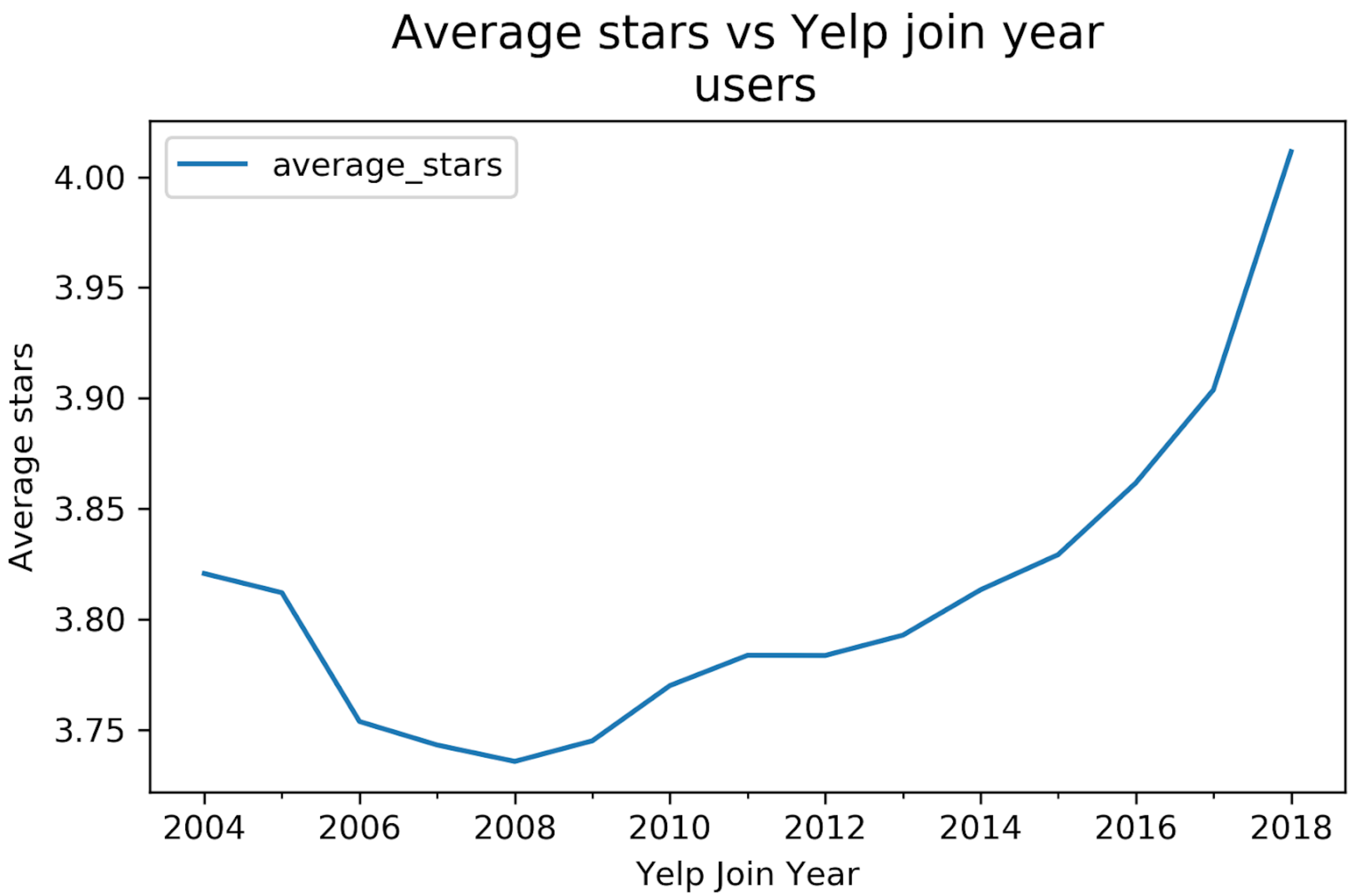
- Highly right-skewed
- 8 people who left even more than 6000 reviews

Monthly New User Counts



- Monthly new users increased over time and then started to decrease after the peak around 2014

Average Stars vs. Join Year



- The users who joined Yelp later tend to have higher average stars (ignoring the first couple years)

Procedures

- Data collecting and wrangling
- Exploratory data analysis (EDA)
  - Businesses
  - Reviews
  - Users
- **Recommender system models**
- Final recommendation

Data Preparation

- Selected Las Vegas–Henderson–Paradise metropolitan area (5 cities) only and cleaned city names
- Selected businesses and users with enough reviews
- Data left
  - 493,658 reviews
  - 20,340 users
  - 6,266 businesses
- Splitted test and training sets (10% vs. 90%)

Building recommender systems

- Predicted star ratings using collaborative filtering and content-based filtering algorithms
- Collaborative filtering
  - 4 algorithms
  - Used Surprise, a Python package developed for recommender systems
- Content-based filtering
  - 3 algorithms
  - Built using the basic concept of content-based filtering
  - Utilized Sci-kit learn
- Evaluated every model using RMSE

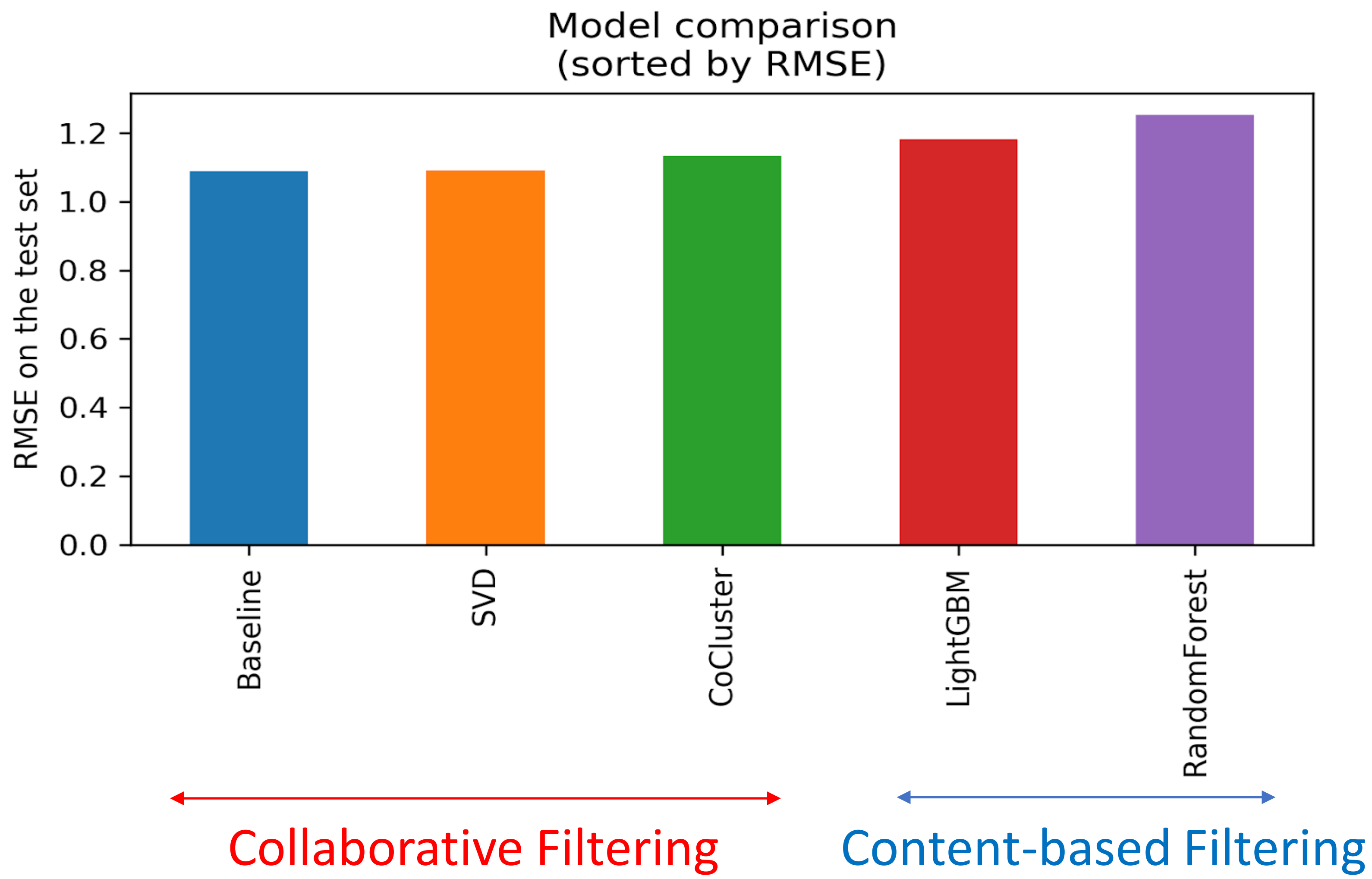
Collaborative Filtering

- Predicted ratings of a user on an item using ratings of other users
- Did not utilize metadata of items or users (content-based filtering)
- Grid search for hyperparameter tuning (if applicable)
- 3 fold cross-validation
- Algorithms
  - Normal Predictor
  - Baseline
  - SVD
  - Co-clustering

Content-based Filtering

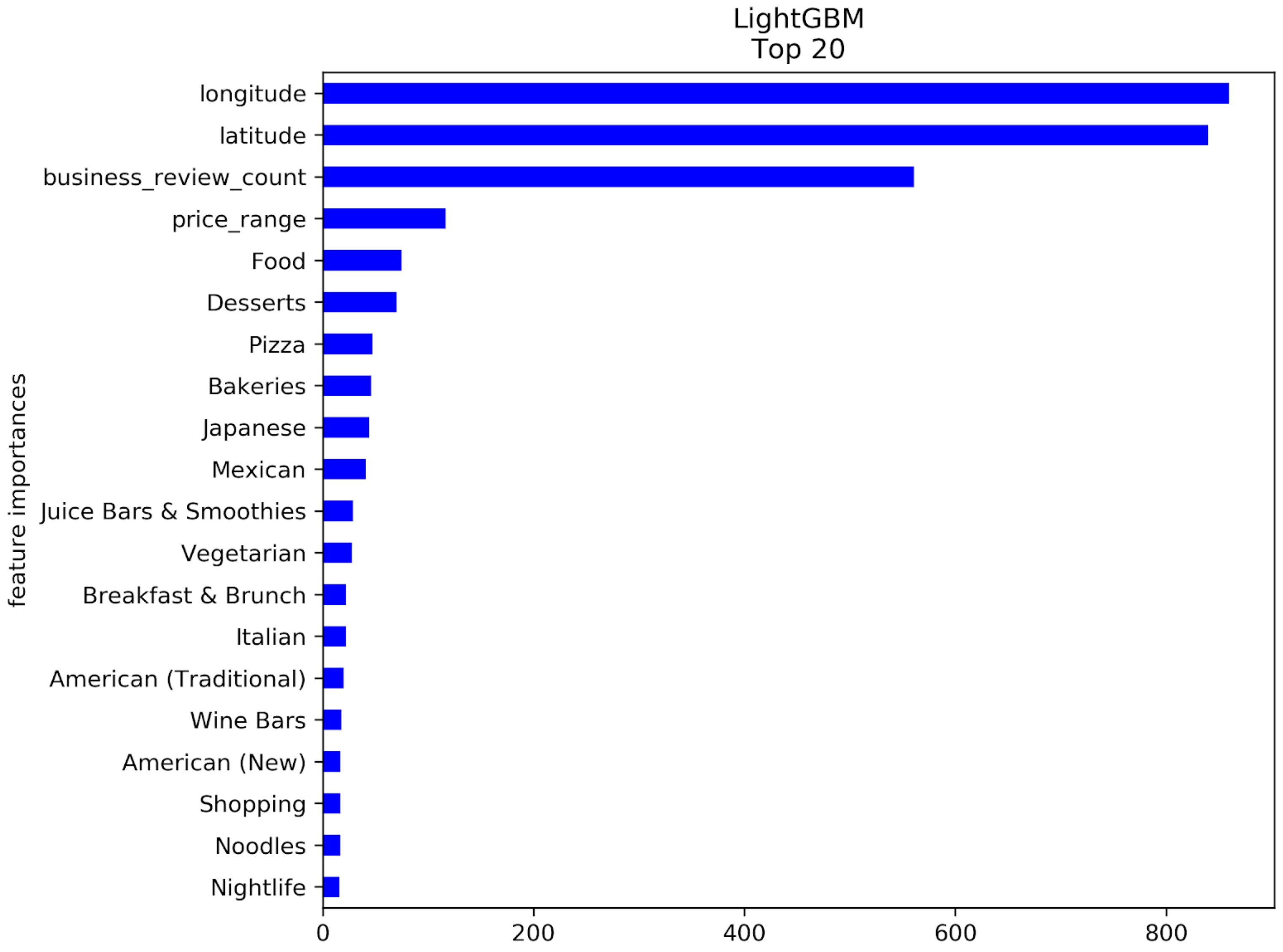
- Built content-based filtering models using the [basic concept of content-based filtering]
- Regression algorithms with the business features as predictors
- Optimized model parameters for each user (for different user tastes)
- Preprocessing
  - Business features: city name, latitude, longitude, prince range, review count and categories (each business had multiple categories, 439 possible)
  - One-hot encoding for 5 kinds of city names
  - Reduced 439 subcategories to 100 using frequency and made one column for each category with 0's and 1's
- Regression algorithms used: Ridge, Random Forest and LightGBM

Top 5 models



- Collaborative filtering models outperformed content-based models

Feature Importances for the top user



- Only 37 nonzero importance features for the top user (LightGBM model).
- The top two features are latitude and longitude

Future Directions

- Error Analysis: the top user has only RMSE of 0.7629 for the LightGBM model. I would like to further investigate users and businesses with big RMSEs.
- The content-based filtering models tried here were made only using the basic concept. I would like to try more complex models for content-based filtering.
- Context-aware collaborative filtering (hybrid of content-based and collaborative filtering) would give the best ressult.

Procedures

- Data collecting and wrangling
- Exploratory data analysis (EDA)
  - Businesses
  - Reviews
  - Users
- Recommender system models
- **Final recommendation**

Final Recommendation from EDA

- Localized marketing strategy
  - EDA showed difference between cities
- Seasonal marketing
  - People leave more reviews during summer and less during winter
  - Summer could be the best season to raise profit for food delivery services
- Preference change over time
  - Review pattern change over time suggests to apply updated rating standards and consider ratings in recent years more
- Target elite users
  - Review counts of users are highly right-skewed and some people left even several thousands of reviews

Final Recommendation from recommender systems

- In general, collaborative filtering would perform better than content-based filtering when recommending restaurants that users might like.
- However, collaborative filtering is not applicable for new businesses or new users (cold start problem) and for such a case, content-based filtering algorithms could come into play.
- Collaborative filtering models also had low error when predictions are made for ratings of users with lots of reviews
- Context-aware collaborative filtering (hybrid of content-based and collaborative filtering) could be the best for better recommendations and new businesses and users.

Links

- Final report

https://docs.google.com/document/d/1_t4QVG-hCVr2dN7XrMNbW1KWa2EkonQcrVeO3WPMRoU/edit?usp=sharing

- Jupyter notebooks on Github

https://github.com/math470/Springboard_Capstone_Project_2